Comparing companies and strategies: a genetic algorithms approach

W.H.G.J. Hennen

LEI (Agricultural Economics Research Institute), Alexanderveld 5, 2585 DB The Hague, the Netherlands; wil.hennen@wur.nl

Abstract

To make good managerial decisions, essential information is needed to discover the causes of a company's good or bad performance. In this perspective, companies should be compared by means of suitable reference values. Face-IT has been developed to compare a company with the average of a comparable group of companies. The entrepreneur or his advisor/accountant can make such external comparison based on the desired set of classification factors, and explore consequences of company's strategies. An application in the dairy sector with accounting data is successfully used in several projects. To find a suitable or even optimal group of companies out of numerous different combinations, a genetic algorithm was developed and implemented as a search mechanism. The objective is to find companies that are comparable, both individual and as a group. This paper goes more deeply into the method of genetic algorithms and how this method is applied in the Face-IT tool. A dairy farm application is used as an example for illustration. As a generic comparison tool, Face-IT can also be applied in surveys, or as a marketing instrument or for exploring success factors.

Keywords: comparison, accounting data, strategic management

Introduction

The availability of appropriate accounting data is an essential condition for management support. Once these data are made available, they can be analysed. Analysis yields management information for (strategic) decision-making. Strong and weak aspects of the entrepreneur's management are deduced as a basis of SWOT[1] analysis for marketing strategies. This provides information on the causes of good or bad performance (Heinrich and Walter, 1989), or for current or future actions (Davis and Olson, 1985). In order to do so, both the data and the evaluation process should be of adequate quality. This paper will focus on one aspect of this evaluation process: the development of suitable reference values for comparing companies.
Different types of reference values or standards can be compared (Hennen, 1995), e.g. data from previous years (historical), data based on planning or simulation models, normative standards based on scientific research, regression models and average performance of a comparable group (external comparison within the same year). The last type is appealing since the entrepreneur is eager to compare his result with similar companies. For such comparison, a method and tool based on genetic algorithms (GA; see e.g. Mitchell, 1998) is developed and described in this paper. This tool, Face-IT[2], searches the optimum combination of companies from numerous possibilities. An entrepreneur who wishes to make his own comparison from a desired set of classification factors or wish to compare different strategies, can compare his own outcome with the average outcome of comparable companies.

[1] Strengths, Weaknesses, Opportunities, and Threats.
[2] Farm Accounts Compared by Evolutional Improving to Top-combination.

The focus of this paper is the description of the (method of the) Face-IT tool. More specifically, the data used, the method of GAs and the way in which algorithms are developed and applied in the Face-IT tool. The functionality is illustrated by an application in the dairy sector and the way in which applications can be developed.

Material and methods

Datasets

Datasets from various domains have potential for a Face-IT application, so different applications with the same tool can be developed. In this paper, a dataset from the Farm Accountancy Data Network (FADN, 2009) with data from nearly 400 Dutch dairy farms was used as an example for Face-IT. Records in this dataset represent farms with characteristics as attributes, e.g. number of cattle, region, milk yield, feeding costs or gross margin. A definition dataset describes the parameters of Face-IT (described below), and the description and range values for each attribute. Both dataset and definition dataset are plain text files (ASCII) or incorporated in a Microsoft Access database. For a proper use of Face-IT, datasets must consist of a large number of records, the more the better. A reasonably homogeneous dataset, i.e. a dataset with a limited number of extreme records (farms), must contain at least a few hundred records. Heterogeneous datasets should contain a few thousand records or more. The dairy dataset used in this paper is not heterogeneous; 400 records are assumed to be enough for an application.

Method of Face-IT

The objective of Face-IT is to find a group of records (or farms) R^o that matches a situation (or farm) of interest (x) as closely as possible, based on the selected attributes (A^s) given by the user. Not only should an individual record $r \in R^o$ have a good match with x, but the average value of each attribute $a \in A^s$ of all records in R^o must also resemble the comparable attributes of x 'as well as possible'. GA was implemented as a search mechanism for this specific problem.

Genetic algorithms as applied in the tool

GA, which was made popular by Holland (1975), is a computer algorithm that solves difficult problems through the principles of evolution. Problems are first translated (encoded) to strings of binary digits called chromosomes. Through the processes of reproduction and natural selection, chromosomes that encode the most successful structures appear and evolve in the population (Davis, 1991). Each chromosome is evaluated and results in what Davis calls 'a measure of the chromosome's performance on the problem to be solved'. Biological mechanisms such as crossover and mutation take place during reproduction where the population of chromosomes gradually improves: in each cycle or generation children with better evaluation results replace parents. Eventually a (local) optimum is reached for the best individual in the population. Although different researchers generally use different implementations of the algorithm, a necessary condition should be the existence of the crossover mechanism (Davis, 1991). In Face-IT, a GA is implemented. Before the programme calls the GA, a pre-selection of the records takes place as the first step of the calculations.

1. *Pre-selection (without GA).* Face-IT has to find a group of records (i.e. farms); in the rest of this paper we will assume that the size of the group is set to 10. Individual members as well as group averages should sufficiently match the farm; this is calculated by weighted differences of criteria or attributes. Finding the best 10 out of 400 farms is an inconceivable problem due to the enormous number of combinations ($400! / (10! * 390!) = 10^{22}$). A practical solution is a

pre-selection based on individual matching, e.g. 10% of the best matching farms (i.e. the user-defined pre-selection pressure λ). After pre-selection, the number of combinations is sharply reduced to less than a billion.

The pre-selection mechanism is formally described as follows. Let R be the collections of records in a database and A be the set of attributes $a_1 \ldots a_j$. Each individual record $r \in R$ will be compared with the particular farm x ($x \in R \wedge x \neq r$) based on a subset of chosen attributes $A^s \subseteq A^p \subseteq A$, where $a_i \in A^s$ and $i \geq 2$. Only an attribute from the set of potential attributes A^p is eligible for selection. There may be attributes in the set A where selection by a user is not allowed. These attributes are merely used for comparison purposes.

After each individual record $r \in R$ is compared or matched by means of a fitness function (Euclidian distance, see below); these records are sorted based on similarity. After sorting, the set R will become R^s. Only $R^p \subset R^s \equiv R$, or the best matching records, will be used in the GA process. The number of records in R^p is calculated as $|R^p| = \lambda * |R^s|$. The pre-selection pressure λ should be small enough to guarantee that all members of R^p resemble our farm x to an acceptable level. Very large databases have opportunities for a large $|A^s|$ and a very small λ, and consequently a very good set R^p. A suitable value for λ depends on the characteristics of the dataset. In order to limit the number of combinations, only a few dozens of records should be in set R^p.

2. *Fitness function*. For matching, a fitness (or evaluation) function is applied. This function is based on the Euclidian distances of the values of $a_i \in A^s$. For a_i, the deviation between r and x is normalised by a pre-defined range (declared in the definition-dataset). The fitness function can also account for user preferences: importance of a_i out of the set {'very important',…, 'slightly important'}, and the maximum allowable variation of the group values of a_i. The importance is translated to a number that gives the power of the standardised difference. Attributes that are more important have a higher value.
3. *Genetic algorithm in Face-IT*. With pre-selection we have assured that each individual farm $r \in R^p$ has a decent match with x. From now on, all farms in the set R^p are treated equally; individual fitness is no longer an issue. Now we want to ensure that the average value of each attribute $a_i \in A^s$ of all farms in $R^o \subset R^p$ resembles the comparable attribute of x 'as well as possible'. The GA is applied to find the 'optimal' set R^o where the weighted average of each chosen attribute in set set A^s matches each attribute of x.

First a random population of individuals is generated in Face-IT, each having a distinct set of genes (i.e. records) on a chromosome (i.e. group of 10 records or farms). A new generation is produced by the Darwinian principle of natural selection or the survival of the fittest. Parents or individuals with the best evaluation of the fitness function are randomly allowed to produce offspring by using the operations *crossover* (swapping genes from the parent's chromosomes) and *mutation* (replacing a gene with a new one from the set R^p).

Better parents are likely to produce more offspring. Because of performance constraints, the time-consuming selection with a roulette wheel (Davis, 1991) was not implemented. Instead, a probabilistic mathematical function was used. The probability of being selected increases linearly with the fitness of the parent. After a new generation with these parents is produced, the generated population is sorted and new parents are selected for the next generation. This continues until the best solution found so far no longer produces any improvement.

Optimisation algorithms, and thus the GA in Face-IT, try to find the optimum within a large number of combinations. They do not guarantee that the optimum will be found, although a solution might be good enough for practical use. To partially solve the problem of 'stuck in local minima', Face-IT starts a number of different runs (e.g. 10 or 20) each with a different population to start with. The best outcome of all runs will be considered as the (sub)optimal solution: R^o. If more runs result in the same best solution, R^o may even be considered optimal. It is not unusual to start different runs with different random seeds to reach better results in a more efficient way (Davis, 1991).

Functionality of the tool

The method of Face-IT is implemented in (1) the visual development environment of Borland Delphi, (2) as an internet system in Visual Basic .Net, and (3) in Excel VBA. The user can operate the tool in a user-friendly way and an experienced user may adjust the settings of the GA. The settings he is allowed to change are: the number of runs and generations, the size of the group (R^o, minimum of 2), the number of individuals in a population, the fraction of parents, the pre-selection pressure λ, and the chance a mutation takes place.

After the GA settings are defined, the record x (e.g. a farm) is selected. The values of all attributes A for x are shown. From the subset $A^p \subseteq A$, the user can choose those attributes or criteria he wants to compare (the selected set $A^s \subseteq A^p$), together with their importance. A restriction of the standard deviation is optional. After pressing the *run* button, the results are presented within a few seconds. Table 1 shows for a particular dairy farm x, the values for some criteria. Column A^s indicates the criteria the user has chosen from the list of potential criteria A^p. This can be done very easily by mouse click, as well as selecting the importance of a criterion from a set of pre-determined set of importance classes. The criterion 'Milk yield per cow' is considered very important and should have a restricted variation ('Max std.') in the solution R^o.

The aim of Face-IT is to find a group of farms with comparable values for the 4 selected criteria, considering the importance and the restriction of the variation. The column 'Average group' shows the results of the comparison group. These results are unweighted averages. The members of the group shown in the 'Comparison group' are individual comparable (due to 10% pre-selection). When we compare 'Value for x' and 'Average group', it can be concluded that the group is fairly similar considering the 4 selected criteria. Especially 'Milk yield per cow', because the user has indicated this as a very important criterion. Besides the 4 selected criteria, we can compare other variables as well. Farm x uses more nitrogen and concentrates compared to similar farms. This may be a good starting point for finding causes and justifications of higher inputs.

Table 1. Illustrative example of Face-IT for a dataset of dairy farms (see text for explanation).

Criterion	A^p	A^s	Value for x	Importance	Max std.	Average group
East-West position	x		2			1.95
North-South position	x		2			1.50
Milk quota / farm	x	x	450,000	Average	-	450,396
Milk yield / cow	x	x	7,800	Very important	500	7,801 (std: 460)
Acreage grassland	x	x	36.5	Average	-	36.0
Acreage maize land	x	x	0	Average	-	0
…						
Nitrogen kg/ha			296			234
Concentrates kg/cow			2,300			2,027
Comparison group	343, 4, 60, 298, 45, 122, 107, 368, 59, 72					

With the Face-IT tool, the user can explore strategic decisions on his farm. Let suppose he is interested in the effect of buying quota (200,000) and maize land (10 hectares), see the underlined values in Table 2 (e.g. 650,000 = 400,000 (original value) + 200,000).
Now a new group of farms is found ('Comparison group'). Based on the 4 criteria, the new group is quite similar to the strategy. When Tables 1 and 2 are compared, it is likely that such a strategy may increase inputs: a difference of +38 and +97 kg respectively for nitrogen supply per hectare and concentrates per cow. However, other factors not corrected for (e.g. region and soil type) may be different. The user can correct for some of these factors and repeat the calculation resulting in a new group with new averages.
It should be clear that the comparison group is a subset of the total dataset. If the farm x has extreme characteristics or if the user defines an extreme strategy (e.g. 1,200,000 'Milk quota / farm'), the programme will not be able to find a suitable comparison group. The dataset contains no such farm or only a very few. This is a major limitation of Face-IT. Although a larger dataset might alleviate this problem somewhat, the user should be aware of this shortcoming. However, this is common to all external comparison methods.

Evaluation of the tool
From the applications that have been made, and in particular the dairy farm application, the following characteristics and requirements of Face-IT can be summarized:
- The most advantageous characteristics of Face-IT are the short development time (a few hours), the ease of making comparisons, the fast calculations by making use of the GA and its versatile applications.
- A user can compose comparable groups very easily, very fast and with limited explanation required. The user himself is responsible for choosing the criteria and their importance.
- In principle, the user needs no statistical or model building knowledge (e.g. regression analysis). It is a tool where the user can play around and explore different situations and become familiar with the domain.
- With the same tool, group comparisons and strategy comparisons can be made.
- Group averages are close to the value of chosen attributes when no extreme criteria are considered and when the dataset contains enough records. Extreme situations or innovative farms

Table 2. Illustrative example of strategy comparison with Face-IT (see text for explanation).

Criterion	A^p	A^s	Value for x	Importance	Max std.	Average group
East-West position	x					2.55
North-South position	x					2.00
Milk quota / farm	x	x	*650,000*	Average	-	653,337
Milk yield / cow	x	x	7,800	Very important	500	7,800 (std: 472)
Acreage grassland	x	x	36.5	Average	-	36.6
Acreage maize land	x	x	*10*	Average	-	9.78
…						
Nitrogen kg/ha						272 (+38)
Concentrates kg/cow						2,124 (+97)
Comparison group	89, 9, 6, 377, 16, 191, 200, 71, 346, 313					

with a specific investment cannot be compared. Users should be aware of this shortcoming, especially when it is used as a tool for strategic management.
- Face-IT is not restricted to certain domains and is therefore generic widely applicable.

Alternatives to genetic algorithms
It is of interest to compare GA with other search methods in the field of combinatorial optimisation (Lawler, 1976) that have potential to solve this type of problem. For example, neighbourhood search (e.g. Papadimitriou and Steiglitz, 1982) or simulated annealing (e.g. Aarts and Korst, 1989). In comparisons, sometimes results with a GA are better (e.g. Mann and Smith, 1996), and sometimes worse (e.g. Lahtinen *et al.,* 1996) than simulated annealing. Manikas and Cain (1996) show that the type of problem is a factor for deciding which is best.

Face-IT applications

Each Face-IT application consists of four parts or files: the dataset, the definition dataset, default GA settings and the tool. The *dataset* where each row contains a record and each column represents an attribute. Each element in the dataset must have a numeric value. Missing values are not allowed; the developer of the application should remove the record concerned or impute with another value (e.g. the average). Although categorical (nominal) variables are allowed in the dataset, the user should be aware of these when interpreting the results. Differences between nominal values only give an indication of the 'difference' with the comparison group. In fact there are no problems with binomial (0/1) variables; the result gives the distribution within the group (e.g. 0.8 must be interpreted as: 8 out of 10 has value 1). For each attribute, the name as well as the value of the range is declared in the *definition dataset*. With this range, the difference between the value of farm x and the value of another farm (or average of the group) is standardised. Sometimes one range is too rude and therefore not enough for fine-tuning. The developer of the application may then declare additional range parameters for different intervals. *Default settings* of the GA parameters are declared, such as the number of runs or the size of the group. The user can always adjust these settings in the Face-IT tool, the fourth part of an application. This is an executable programme, see the previous section for its functionality.
The development of a new application from an existing dataset can be completed within a day, where the main focus should be on the definition dataset, especially on the range parameters. Face-IT is a generic tool and therefore domain-independent. Each dataset is suitable as long as this dataset has the required format, contains a large number of records, is complete (no missing values), has no non-numeric values and is not dominated by nominal attributes. The generic applicability of Face-IT is illustrated in Table 3. As case serves an example supplied by the SPSS package. The sample file 'bankloan.sav' consists of 700 records, characteristics of people who are likely or not to default on loans (good and bad credit risks). In this set, about 26% of the people have defaulted. Suppose we want to compare person x with a group of 10 persons with the same age, years employed, years at current address, income and the same percentage debt. Face-IT finds a group with good resemblance, e.g. the group average for 'Age' is equal to the age of person x (Table 3). Neither person x nor someone from the comparison group does default, but there are considerable differences regarding the other characteristics (e.g. level of 'education'). Suppose person x had a different 'percentage debt', let say 10%. For this new person, named x', a comparable group can be found (last column of Table 3). Again the resemblance is quite good. The effect of the change of 'percentage debt' from 5.8% to 10% becomes clear if averages of both groups are compared. The education level is lower and the number of people who defaulted higher.

Table 3. Illustrative example of strategy comparison with Face-IT for the Bankloan case (see text for explanation).

Criterion	A^p	A^s	Value for x	Average group	Value for x'	Average group
Age	x	x	33	33.0	33	33.0
Education	x		1	1.8		1.3
Years employed	x	x	8	8.1	8	7.9
Years at address	x	x	7	7.1	7	6.8
Income	x	x	27	26.6	27	27.5
Percentage debt	x	x	5.8	5.83	*10*	9.9
Debt credit card	x		0.78	0.33		0.78
Debt other	x		0.79	1.2		1.96
Default			0	0.0 (=0%)		0.30 (=30%)
Comparison group	Consist of 10 persons (i.e. 10 records of bankloan.sav file)					

Experience with Face-IT

Most experience has been obtained with a dairy farming application that is used successfully in some projects (e.g. Beldman and Doornewaard, 2002). Face-IT is applied to find comparison ('mirror') groups. Or to create data for a game simulation model (Beldman and Daatselaar, 2002; Hennen, 1995), when no dataset from a participating dairy farmer is available during the game. Farmers were asked values for a handful of characteristics (e.g. acreage, number of livestock). Face-IT then searched for a comparison group and averaged the results in a set ready for the game simulation model. Some farmers who used such fictitious data in a game simulation session felt as if they were working with their own data.

Agricultural students also make use of this application (Hurkens, 2006; GKC, 2009). During their work placements, students become acquainted with the circulation of knowledge between enterprises, teaching institutes and research centres.

The dairy farming application can be consulted both web-based and locally. Other applications have also been developed (e.g. arable farms). An interesting application is one based on a survey dataset. A group of respondents who give similar answers to certain key questions can be extracted.

Conclusion

This paper presents a method and tool based on GA aimed at finding a group of records that matches a record, situation or strategy as closely as possible, based on the values of selected attributes chosen by the user. Not only are individual group members comparable (due to pre-selection), but also the group averages of the selected criteria. Farm results can be mirrored in order to compare performance and help the farmer to analyse strong and weak points (Beldman *et al.,* 2006).

With two cases, the dairy farming case and the bank loan case, this paper demonstrates that GA is a suitable technique for such a problem. For a dairy farmer it is possible to make his preferred comparison group easily without any statistical inconvenience, and as a result, he has the ability to compare his farm with similar farms.

Some alternatives to GA are mentioned in this paper. Simulated annealing has been implemented as well in the Excel/VBA version of Face-IT. Applied with the same case (dataset of dairy farms), the method of GA is faster than simulated annealing and gives better results (W.H.G.J. Hennen, personal communications).
Face-IT has potential for teaching purposes and is still being used in schools. But there are other recommended potential applications. Accounting firms might find it interesting to add comparison figures in business reports; a researcher could find comparable respondents in a survey and compare groups, a CEO of a large organisation could explore success factors of management teams without feeling inhibited by a possible lack of statistical knowledge; a market analyst might discover interesting facts about customer behaviour based on information of purchases and bonus cards, etc.

References

Aarts E., and Korst J., 1989. Simulated Annealing and Boltzmann Machines. New York, John Wiley & Sons.

Beldman A.C.G., and Daatselaar C.H.G., 2002. Focus on the entrepreneur, use of Game Simulation Dairy in nutrient management. Proceedings VIAS symposium. Agro-informaticareeks 13.

Beldman A.C.G., and Doornewaard G.J., 2002. Meer zicht op de cijfers, analyse technische en economische resultaten K&K-bedrijven in vergelijking met andere praktijkbedrijven. Koeien & Kansen rapport 8, Praktijkonderzoek Veehouderij, Lelystad.

Beldman A.C.G., Hennen W.H.G.J., and Doornewaard G.J., 2006. The Use of a Benchmark Tool Based on FADN for the Management of Organic Dairy Farms. In: Rippin E. *et al.* (eds) Towards a European Framework for Organic Market Information. Proceedings of the Second EISfOM European Seminar, Brussels, November 10 & 11, 2005.

Davis L., 1991. A genetic algorithms tutorial. In: Davis L. (ed.) Handbook of Genetic Algorithms. Van Nostrand Reinhold, 1991, 1-101.

Davis G.B., and Olson M.H., 1985. Management information systems: Conceptual foundations, structure, and development. New York: McGraw-Hill.

FADN, 2009. Available at: http://ec.europa.eu/agriculture/rica Assessed 5 January 2009.

GKC, 2009. Available at: www.groenekenniscooperatie.nl Assessed 5 January 2009.

Heinrich I., and Walter K., 1989. Kontrolle and Analyse des Betriebsweiges Futterbau – Rindviehaltung mit Personalcomputer. Agrarwirtschaft 38(5): 135-150.

Hennen W.H.G.J., 1995. Detector: knowledge-based systems for dairy farm management support and policy analysis; methods and applications. PhD thesis. LEI, The Hague.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor.

Hurkens R., 2006. Ondernemerschap in het Groen Onderwijs via AgroCenter en hoe verder ...: Groen onderwijs aan het roer: een startpunt voor een nieuwe ontwikkeling. LEI, The Hague (in Dutch).

Lahtinen J., Myllymäki P., Silander T., and Tirri H., 1996. Empirical comparison of stochastic algorithms in a graph optimization problem. Proceedings of the Second Nordic Workshop on Genetic Algorithms, Vaasa, Finland.

Lawler, E., 1976. Combinatorial Optimization: Networks and Matroids. Holt, Rinehart, Winston, New York.

Manikas T, and Cain J., 1996. Genetic Algorithms vs. Simulated Annealing: A. Comparison of Approaches for Solving the Circuit Partitioning Problem. Technical Report: 96-101, University of Pittsburg.

Mann J.W., and Smith G.D., 1996. A Comparison of Heuristics for Telecommunications Traffic Routing. In: Rayward-Smith *et al.* (eds) Modern Heuristic Search Methods. John Wiley & Sons.

Mitchell M., 1998. Introduction to Genetic Algorithms. MIT Press.

Papadimitriou C.H., and Steiglitz K., 1982. Combinatorial Optimization: Algorithms and Complexity. Prentice-Hall, New York.